

AD-A182 565

MINIMUM DISTANCE ESTIMATION OF MIXTURE PROPORTIONS(U)  
AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOOL  
OF ENGINEERING R N BENTON-SANTO DEC 86

1/1

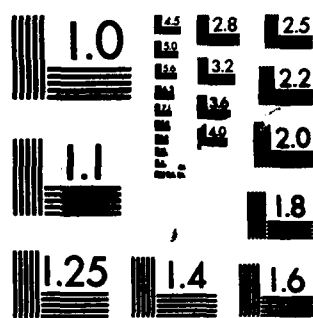
UNCLASSIFIED

AFIT/GOR/MA/86D-1

F/G 12/3

ML

END  
DATE  
FILMED  
8



MICROCOPY RESOLUTION TEST CHART  
NATIONAL BUREAU OF STANDARDS-1963-A

DTIC FILE COPY

1

AD-A182 565



MINIMUM DISTANCE ESTIMATION OF  
MIXTURE PROPORTIONS

THESIS

R. Nicole Benton-Santo  
Second Lieutenant, USAF

AFIT/GOR/MA/86D-1

DTIC  
ELECTE  
JUL 24 1987  
S E D

DEPARTMENT OF THE AIR FORCE  
AIR UNIVERSITY

**AIR FORCE INSTITUTE OF TECHNOLOGY**

Wright-Patterson Air Force Base, Ohio

This document has been approved  
for public release and sale; its  
distribution is unlimited.

87 7 22 079

AFIT/GOR/MA/86D-1

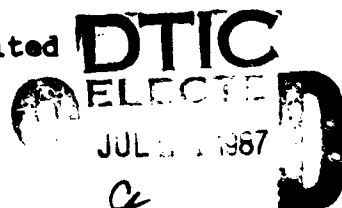
MINIMUM DISTANCE ESTIMATION OF  
MIXTURE PROPORTIONS

THESIS

R. Nicole Benton-Santo  
Second Lieutenant, USAF

AFIT/GOR/MA/86D-1

Approved for public release; distribution unlimited



AFIT/GOR/MA/86D-1

MINIMUM DISTANCE ESTIMATION OF  
MIXING PROPORTIONS

THESIS

Presented to the Faculty of the School of Engineering  
of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the  
Requirements for the Degree of  
Master of Science

R. Nicole Benton-Santo, B.S. Mth  
Second Lieutenant, USAF

December 1986

|                    |  |
|--------------------|--|
| Accession For      |  |
| NTIS GRA&I         | <input checked="checked" type="checkbox"/> |
| DTIC TAB           | <input type="checkbox"/>                   |
| Unannounced        | <input type="checkbox"/>                   |
| Justification      |  |
| By _____           |  |
| Distribution/      |  |
| Availability Codes |  |
| Dist               | Avail and/or<br>Special                    |
| A-1                |  |

Approved for public release; distribution unlimited



### Preface

The purpose of this study was to compare Minimum Distance estimation techniques to the Method of Moments and quasi-clustering techniques when applied to mixed exponential and mixed normal distributions. A Monte Carlo simulation was performed, using the Anderson-Darling goodness-of-fit statistic as a distance measure for the Minimum Distance method. Finally, the mean-square error was used as the performance criteria in the comparison step.

I wish to thank my thesis adviser, Dr. Albert H. Moore, for suggesting this topic and also for his constant encouragement and assistance. I am also deeply indebted to 1Lt. Mark Gallagher and Capt. Greg McIntyre for their invaluable assistance in the development of the computer software. A word of thanks is also given to my parents, Randall and Nedra Benton, for their long-distance words of encouragement and support. Finally, I would like to extend my appreciation to my husband, Fred, for typing the thesis and for his undying sacrifice and understanding during the past few months.

R. Nicole Benton-Santo

Table of Contents

|   | <u>Page</u> |
|---|-------------|
| Preface . . . . .   | ii          |
| List of Tables . . . . .  | v           |
| Abstract . . . . .  | vi          |
| I. Introduction . . . . .   | 1           |
| II. Estimation . . . . .  | 3           |
| Criteria of Good Estimators . . . . .                             | 3           |
| III. The Mixed Model . . . . .                                    | 5           |
| Previous Research . . . . .                                       | 6           |
| IV. Estimation Techniques and Mixed Distributions. . . . .        | 8           |
| The Method of Moments . . . . .                                   | 8           |
| Moments of the Mixed Normal . . . . .                             | 9           |
| Moments of the Mixed Exponential . . . . .                        | 11          |
| The Method of Maximum Likelihood . . . . .                        | 13          |
| Maximum Likelihood and the Mixed Normal . . . . .                 | 14          |
| Maximum Likelihood and the Mixed Exponential . . . . .            | 15          |
| V. Minimum Distance Estimation . . . . .                          | 17          |
| Kolmogorov-Smirnov . . . . .                                      | 18          |
| Cramer-von Mises . . . . .  | 19          |
| Anderson-Darling . . . . .  | 19          |
| VI. Methodology . . . . .   | 22          |
| Monte Carlo Simulation . . . . .                                  | 22          |
| Sample Generation from the Mixed Normal . . . . .                 | 23          |
| Sample Generation from the Mixed Exponential . . . . .            | 24          |
| Initial Parameter Estimates for the Normal Mixture . . . . .      | 25          |
| Initial Parameter Estimates for the Exponential Mixture . . . . . | 27          |
| Minimum Distance Estimates . . . . .                              | 27          |
| Comparison Criteria . . . . .                                     | 29          |
| VII. Conclusions and Recommendations . . . . .                    | 31          |
| Appendix : Computer Code . . . . .                                | 35          |

|                        | <u>Page</u> |
|------------------------|-------------|
| Bibliography . . . . . | 40          |
| Vita . . . . .         | 43          |



| Table | <u>List of Tables</u>                                    | <u>Page</u> |
|-------|--|-------------|
| I.    | Simulation Results for Mixtures of Two Exponentials      | 33          |
| II.   | Simulation Results for Mixtures of Two Normal Components | 34          |

Abstract

Minimum Distance estimation was used to calculate estimates of the mixing proportion of the mixture of two normal distributions and the mixture of two exponential distributions. The estimation was carried out by using the Golden Search technique to minimize the Anderson-Darling goodness-of-fit statistic. A Monte Carlo simulation was run for both distribution mixtures, varying the mixing proportions from .25, .5 to .75 with sample sizes of 100 for the normal mixture and 750 for the mixture of exponentials. The simulation was run 500 times for each parameter combination.

An ad hoc quasi-clustering technique was used to obtain the initial estimates for the parameters of the mixed normal while the method of moments technique was used to obtain initial estimates for the mixed exponential parameters. These estimates were then used to start the minimum distance routines which were used to obtain new estimates of the mixing proportions.

Finally, the mean square errors were calculated for use as a means of comparison for the different estimation procedures.

## MINIMUM DISTANCE ESTIMATION OF A MIXING PROPORTION

### I. Introduction

In recent years, there has been an increasing interest in system reliability by the military. This interest has been spurred by the higher cost of new technology weapon systems. By the use of statistics, a system's ability to perform a given mission and its reliability can be predicted. This prediction and related predictions can be used to procure superior systems by the use of multi-criteria decision analysis or as an aide to improving reliability through engineering design. These are only a few examples of the many uses of reliability theory by the military, however, the first step in the process is reliability parameter estimation.

Every system's time-to-failure can be determined by a probability density function (p.d.f.) and its associated cumulative distribution function (c.d.f.) or by a mixture of p.d.f.'s and their associated c.d.f.'s. Classically, the functional form of this c.d.f. is hypothesized by one of the classical distributions, based on a set of test data. After the functional form has been hypothesized, its associated

parameters must be estimated. Prediction of system failure rates depends largely on how well these parameters are estimated using the test data. Many methods have been used for this estimation, for example, the method of moments, linear maps, and maximum likelihood estimation (MLE). All of these methods are widely accepted and are well documented for common distributions and a few distribution mixtures. The final step in the process is to perform a goodness-of-fit test which measures the distance between the estimated distribution and the hypothesized sample distribution function.

Recently, a new technique known as minimum distance estimation, has been developed which, in a sense, "reverses" the above procedure. The functional form of the distribution is still hypothesized, however, the parameters of the p.d.f. are chosen so as to minimize the goodness-of-fit statistic. This procedure has been applied to most of the classical distributions with good success. However, the process has not been widely used when investigating mixed distributions.

The objective of this research is to compare the initial parameter estimates obtained from the method of moments and quasi-clustering techniques to those obtained by the use of minimum distance estimation. In doing this, the search for a better estimation technique will be advanced.

## II. Estimation

When the military uses a probability density function to approximate or describe a system's reliability, many of the parameters associated with the p.d.f. are unknown and must be estimated. There are several widely accepted methods used for this estimation, such as the Method of Moments, the Method of Maximum Likelihood, the Method of Least Squares and Bayesian Estimation. Of these, the Method of Moments and Maximum Likelihood have been the most widely used. Recently, a new method has come under investigation. This new method is known as Minimum Distance Estimation. With all of these different methods of estimation available the question of which method provides the best parameter estimates arises.

### Criteria of Good Estimators

Basically, there are four criteria that are used either individually or collectively to define good estimators.

These criteria are

1. Invariance and Unbiasedness
2. Consistency
3. Efficiency
4. Minimum risk invariant and minimum variance unbiased

(29:73)

It is up to the user to decide which of the above criteria is most important, however, the estimation methods listed previously all produce estimates which fit at least some of the above criteria. Following chapters will contain explanations of the most commonly used methods, and the newest method, when applied to mixed distributions.

### III. The Mixed Model

As was stated in the introduction, any system's time-to-failure can be determined by a p.d.f. and its corresponding c.d.f. or by some mixture of p.d.f.'s and their corresponding c.d.f.'s. In fact, quite often a system can experience more than one type of failure, for example, fatigue failure or a sudden electrical failure. When this is true, instead of using a single distribution to describe the time-to-failure, the mixed model is applicable. Actually, many problems involving mixtures arise in several different fields of study other than reliability theory, for example, biology (13) and remote sensing and space applications (4,32). This explains the ever increasing interest in parameter estimation for the mixed model.

Mathematically, a k-fold mixed c.d.f. is defined as

$$F_X(x) = \sum_{i=1}^n p_i F_{x_i}(x) \quad 0 < p < 1 \text{ and } \sum_{i=1}^n p_i = 1 \quad (3.1)$$

and a k-fold mixed p.d.f. is defined as

$$f_X(x) = \sum_{i=1}^n p_i f_{x_i}(x) \quad (3.2)$$

where

$F_{x_i}(x)$  = the  $i$ th subpopulation c.d.f.

$f_{x_i}(x)$  = the  $i$ th subpopulation p.d.f.

$0 < p < 1$  = the mixing proportion parameter

It has been common practice to assume that the distributions in a mixture belong to the same family, but differ in parameter values.

### Previous Research

With a few exceptions most of the previous research involving estimation of the parameters of mixed distributions involves the Method of Moments, Maximum Likelihood and Bayesian techniques, all of which have been proven to be difficult. Such researchers as Hosmer(12,21,22), Cohen(9), Tan and Chang(7), Hasselblad(16), Hill(18), Day(11), Bowdon and Dick(6), and James(25) have all considered the problem of estimating the parameters in mixed normals. These are only a few of the authors of extensive literature involving the mixed normal. This extensiveness is due to the diverse applications of the mixed normal.

The second most widely explored mixture is that of the exponential. Cohen(9), Hill(18), Hasselblad(17), Rider(34), Boardman(5), Ashour and Jones(2) and Hader and Mendenhall(15) have all researched the area. A few researchers have extended this research to the mixed Weibull, Falls(14) and Rider(35) using the method of moments and Kao(26) using a graphical method, for example.

In all of this research nearly all of the authors are in agreement about the difficulty of the estimation problem, due to the complexity of the equations. In the following chapters, the Method of Moments, the Method of Maximum Like-



likelihood and Minimum Distance estimation will be briefly developed, including the required equations.

#### IV. Estimation Techniques and Mixed Distributions

##### The Method of Moments

One of the oldest and simplest methods for deriving point estimators is the Method of Moments. Due to its lack of sophistication, this method has its faults. For example, the estimators are usually not unbiased or the most efficient. However, on the positive side, the method is relatively easy to employ, it provides consistent estimators, and most importantly, it often provides estimators when other methods fail to do so. (30:360)

The  $k$ th moment about the origin of a random variable is defined as

$$\mu'_k = E(Y^k) \quad (4.1)$$

and can be found by taking the  $k$ th derivative of the corresponding moment generating function if it exists, and evaluating it at  $t=0$ . Mathematically, we can write

$$\mu'_k = d^k m(t)/dt^k \Big|_{t=0} \quad (4.2)$$

The corresponding  $k$ th sample moment is the average

$$m'_k = 1/n \sum_{i=1}^n y_i^k \quad (4.3)$$

The method of moments assumes that these sample moments will be good estimates of the corresponding population moments and that the needed population moments exist. Since the population moments are the functions of the population parameters, the corresponding equations can be equated and solved for the desired parameters. In other words, "Choose as estimates those values of the parameters that are solutions of the equations  $\mu'_k = m'_k$ ,  $k=1,2,\dots,t$ , where  $t$  equals the number of parameters." (30:357)

#### Moments of the Mixed Normal

In this research, a mixture of two normally distributed populations is being examined. The normal cumulative distribution is defined as

$$F(y) = \int_{-\infty}^y (2\pi\sigma^2)^{-1/2} \exp -1/2 (y-\mu/\sigma)^2 dy \quad (4.4)$$

and the normal probability density function is defined as

$$f(y) = (\exp -(y-\mu)^2/2\sigma^2) (2\sigma\pi)^{-1/2} \quad (4.5)$$

where  $\sigma > 0$ , is the standard deviation or the scale parameter

-  $\infty < u < \infty$ , is the mean or the location parameter

-  $\infty < y < \infty$

Therefore, the mixed c.d.f. will be

$$F(x) = p \int_{-\infty}^x (2\pi\sigma_1^2)^{-1/2} \exp -1/2(y-\mu_1/\sigma_1)^2 dy \\ + (1-p) (2\pi\sigma_2^2)^{-1/2} \exp -1/2(y-\mu_2/\sigma_2)^2 dy \quad (4.6)$$

From this c.d.f. it can be seen that there are a total of five parameters ( $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p$ ) that must be estimated. Therefore, the first five moments must be derived.

The moment generating function for the normal distribution is

$$\exp(\mu t + (t^2 \sigma^2)/2) \quad (4.7)$$

Taking the first five derivatives and evaluating at  $t=0$ , we get the following moments:

$$(\mu + \sigma^2 t) \exp(\mu t + (t^2 \sigma^2)/2) \Big|_{t=0} = \mu \quad (4.8)$$

$$\sigma^2 \exp(\mu t + (t^2 \sigma^2)/2) + (\mu + \sigma^2 t)^2 \exp(\mu t + (t^2 \sigma^2)/2) \Big|_{t=0} \\ = \mu^2 + \sigma^2 \quad (4.9)$$

$$\sigma^2 \left[ (\mu + \sigma^2 t) \exp(\mu t + (t^2 \sigma^2)/2) + 2(\mu + \sigma^2 t)(\sigma^2) \exp(\mu t + (t^2 \sigma^2)/2) + (\mu + \sigma^2 t)^3 \exp(\mu t + (t^2 \sigma^2)/2) \right] \Big|_{t=0} \\ = 3\mu\sigma^2 + \sigma^3 \quad (4.10)$$

$$\sigma^2 \left[ \sigma^2 \exp(\mu t + (t^2 \sigma^2)/2) + (\mu + \sigma^2 t)^2 \exp(\mu t + (t^2 \sigma^2)/2) + 2\sigma^4 \exp(\mu t + (t^2 \sigma^2)/2) + 2\sigma^2(\mu + \sigma^2 t)^2 \exp(\mu t + (t^2 \sigma^2)/2) + 3(\mu + \sigma^2 t)^2(\sigma^2) \exp(\mu t + (t^2 \sigma^2)/2) + (\mu + \sigma^2 t)^4 \exp(\mu t + (t^2 \sigma^2)/2) \right] \Big|_{t=0}$$

$$= 3\sigma^4 + 6\sigma^2\mu^2 + \mu^4 \quad (4.11)$$

$$\begin{aligned} & \left[ 2(\mu + \sigma^2 t)(\sigma^4) + 2(\mu + \sigma^2 t)(2\sigma^4 + 6\sigma^4(\mu + \sigma^2 t)) \right. \\ & \quad \left. + 4(\mu + \sigma^2 t)^3(\sigma^2) \right] \exp(\mu t + (t^2 \sigma^2)/2) \\ & \quad + (\mu + \sigma^2 t) \exp(\mu t + (t^2 \sigma^2)/2) \left[ \sigma^4 + \sigma^2(\mu + \sigma^2 t)^2 \right. \\ & \quad \left. + 2\sigma^4 + 2\sigma^2(\mu + \sigma^2 t)^2 + 3\sigma^2(\mu + \sigma^2 t)^2 + (\mu + \sigma^2 t)^4 \right] \Big|_{t=0} \\ & = \mu^5 + 10\mu^3\sigma^2 + 15\mu\sigma^4 \quad (4.12) \end{aligned}$$

These five moments, when equated to their corresponding sample moments and rearranged to accommodate the mixture, yield the following equations which, when solved, provide the five parameter estimates.

$$\sum x_1/n = p\mu_1 + (1-p)\mu_2 \quad (4.13)$$

$$\sum x_1^2/n = p(\mu_1^2 + \sigma_1^2) + (1-p)(\mu_2^2 + \sigma_2^2) \quad (4.14)$$

$$\sum x_1^3/n = p(3\mu_1\sigma_1^2 + \mu_1^3) + (1-p)(3\mu_2\sigma_2^2 + \mu_2^3) \quad (4.15)$$

$$\begin{aligned} \sum x_1^4/n = & p(3\sigma_1^4 + 6\sigma_1^2\mu_1^2 + \mu_1^4) + (1-p)(3\sigma_2^4 \\ & + 6\sigma_2^2\mu_2^2 + \mu_2^3) \quad (4.16) \end{aligned}$$

$$\begin{aligned} \sum x_1^5/n = & p(\mu_1^5 + 10\mu_1^3\sigma_1^2 + 15\mu_1\sigma_1^4) + (1-p)(\mu_2^5 \\ & + 10\mu_2^3\sigma_2^2 + 15\mu_2\sigma_2^4) \quad (4.17) \end{aligned}$$

#### Moments of the Mixed Exponential

The second distribution that will be used in this

research is the exponential. The exponential c.d.f. and p.d.f. are respectively,

$$F(t) = 1 - \exp(-t/\theta) \quad t > 0 \quad (4.18)$$

$$f(t) = (1/\theta) \exp(-t/\theta), \quad t > 0, \quad \theta > 0 \quad (4.19)$$

From the above we get the mixed c.d.f. of

$$F(x) = p(1 - \exp(-x/\theta_1)) + (1-p)(1 - \exp(-x/\theta_2)) \quad (4.20)$$

As with the normal, we can see that there are three parameters ( $\theta_1, \theta_2, p$ ) that must be estimated. This requires the first three moments which can be derived from the exponential moment generating function. This function is

$$M_x(t) = (1 - \theta t)^{-1} \quad (4.21)$$

As before, we take the first three derivatives and evaluate at  $t=0$  to yield the following,

$$1. \quad \theta(1 - \theta t)^{-2} \Big|_{t=0} = \theta \quad (4.22)$$

$$2. \quad 2\theta^2(1 - \theta t)^{-3} \Big|_{t=0} = 2\theta^2 \quad (4.23)$$

$$3. \quad 6\theta^3(1 - \theta t)^{-4} \Big|_{t=0} = 6\theta^3 \quad (4.24)$$

Once again, equate the above population moments to their corresponding sample moments and rearranged for a mixture to obtain the required equations.

$$\sum x_1/n = p \theta_1 + (1-p) \theta_2 \quad (4.25)$$

$$\sum x_1^2/n = 2p \theta_1^2 + (1-p) 2 \theta_2^2 \quad (4.26)$$

$$\sum x_1^3/n = 6p \theta_1^3 + (1-p) 6 \theta_2^3 \quad (4.27)$$

#### The Method of Maximum Likelihood

The Method of Maximum Likelihood is a much more sophisticated method of estimation than the method of moments. In being so, it is more difficult to apply, however, the the method is usually better than the method of moments and yields asymptotically best estimators which explains its popularity.

The method begins with a random sample from a probability density function with unknown parameters. The likelihood function is now defined as the joint density function of the  $n$  random variables, and can be written as,

$$L = \prod_{i=1}^n f_x(x_i; \theta) \quad (4.28)$$

where  $\theta$  is an unknown parameter

If the random sample is from a multiparameter density function, then

$$L = \prod_{i=1}^n f_x(x_i, \theta_1, \theta_2, \theta_3, \dots, \theta_k) \quad (4.29)$$

The maximum-likelihood estimators (MLE) of the parameters are the values of the parameters which maximize  $L$  or equivalently  $\ln L$ . Quite often these values are the solutions to the following equations.

$$\partial \ln L / \partial \theta_i = 0 \quad i=1, 2, \dots, k \quad (4.30)$$

(29:82)

#### Maximum-Likelihood Estimation and the Mixed Normal

Maximum-likelihood estimation of the mixed normal, in its long history, has proven to be very difficult. This difficulty arises largely in the fact that the likelihood equations do not have an explicit solution and instead, must be solved by iterative methods.

Given

$$f_i = \text{frequency of the observations} \quad (4.31)$$

$$q_{ij} = 1/(2\pi \sigma_j^2)^{1/2} \exp(-(x_i - \mu_j)^2/2\sigma_j^2) \quad (4.32)$$

we can let

$$Q_i = \sum_{j=1}^k q_{ij} p_j \quad (4.33)$$

so that the log of the likelihood function is approximately



$$L = \sum_{i=1}^n f_i \ln(Q_i) \quad (4.34)$$

If we take the partial derivatives and set them equal to zero we get the following set of equations that must be solved iteratively,

$$0 = \partial L / \partial \mu_j = \sum_{i=1}^n f_i / Q_i \quad p_j q_{ij} (x_i - \mu_j) / \sigma_j^2 \quad j=1,2,\dots,k \quad (4.35)$$

$$0 = \partial L / \partial \sigma_j = \sum_{i=1}^n f_i / Q_i \quad p_j q_{ij} (x_i - \mu_j)^2 / \sigma_j^3 \quad j=1,2,\dots,k \quad (4.36)$$

$$0 = \partial L / \partial p_j = \sum_{i=1}^n f_i / Q_i \quad (q_{ij} - q_{ik}) \quad j=1,2,\dots,k-1 \quad (4.37)$$

Two possible methods for solving these equations could be Newton's Method or the method of Steepest Descent.

(16:432-436)

#### Maximum Likelihood Estimation and the Mixed Exponential

Similarly, for the mixed exponential where,

$$f_j(x) = 1/\theta \exp(-x/\theta) \quad \text{for } x > 0 \quad (4.38)$$

and

$$g(x) = \sum_{j=1}^k p_j f_j(x) \quad (4.39)$$

we can write the log of the likelihood function as

$$L = \sum_{i=1}^n \ln g(x_i) \quad (4.40)$$

By taking the partial derivatives, equating them to zero, and simplifying as illustrated by Hasselblad (1966), we get the following equations which must be solved by "successive substitution" iteration. (17:1462)

$$p_1^{(v+1)} = p_j^{v/N} \sum_{i=1}^n f_j(x_k)/g(x_1) \quad (4.41)$$

$$\theta_j^{(v+1)} = \sum_{i=1}^n f_j^{(v)}(x_1)x_1/g^{(v)}(x_1) / \sum_{i=1}^n f_j^{(v)}(x_1)/g^{(v)}(x_1) \quad (4.42)$$

where  $v+1$  = current iteration

$v$  = last iteration

## V. Minimum Distance Estimation

Minimum Distance estimation is a relatively new estimation technique which is in the forefront of statistical research. The method was first developed by J. Wolfowitz in a series of articles beginning in 1952.(36,37) In these articles he developed the method as a means of deriving "super-consistent" estimators when classical methods failed to. (36: 1953). During the following years much work was done in the field of minimum distance applications to classical distributions. Much of this work was done by Dr. Albert Moore and his thesis students. For example, Hobbs, Moore and James(19) investigated minimum distance estimation and the three parameter Gamma distribution while Hobbs, Moore, and Miller(20) applied the method to the three parameter Weibull. Very recently minimum distance estimation has been under study for robustness. Parr and Schucany(33) were two of the first to investigate this area and did show that minimum distance techniques provide robust estimators for the location parameters of symmetric distributions. Since then Daniels(10) applied Parr's methodology to the generalized t distribution, and Miller(31) to the three parameter Weibull, among others.

Wolfowitz's minimum distance method begins by defining the distance between two distribution functions as

$$(F_1, F_2) = \sup_x |F_1(x) - F_2(x)| \quad (5.1)$$

where sup means supremum.

The method is in no way tied to this definition and can be applied with many different definitions of distance. (36:10)  
For the purpose of this research however, the above definition of distance will be used.

The method next selects a random sample of size  $n$  from a known distribution and determines the empirical distribution,  $S_n(x)$ . The empirical distribution is defined as

$$S_n(x) = \text{number of values } < x / \text{total number in sample } (n) \quad (5.2)$$

and is a non-decreasing step function with jumps of size  $1/n$  at each observed value. (3:269)

The final step is to determine the parameter estimates by minimizing the calculated distances between the empirical distribution function and the true or hypothesized distribution. There are several different distance measures that are commonly used for this step, including the Kolmogorov-Smirnov, the Cramer-von Mises, and the Anderson-Darling statistics.

#### Kolmogorov-Smirnov

The Kolmogorov-Smirnov statistic is defined as

$$D = \sup |F(x) - S_n(x)| \quad (5.3)$$

In other words, the largest absolute deviation between the hypothesized distribution and the empirical distribution. Computationally,

$$D = \max(D^+, D^-) \quad (5.4)$$

where  $D^+$  and  $D^-$  are defined as follows

$$D^+ = \max\{(i/n) - z_i\}, \quad 1 \leq i \leq n \quad (5.5)$$

$$D^- = \max\{z_i - (i-1)/n\}, \quad 1 \leq i \leq n \quad (5.6)$$

and

$$z_i = F(x_i) \quad (5.7)$$

(3:269-270)

#### Cramer-von Mises

The Cramer-von Mises distance technique is defined by

$$W^2 = n \int_{-\infty}^{\infty} [S_n(x) - F(x)]^2 \downarrow (F(x)) dF(x) \quad (5.8)$$

where  $(F(x))$  is a weighing factor or computationally

$$W^2 = \left[1/(12n)\right] + \sum_{i=1}^n \left[z_i - (2i-1)/2n\right]^2 \quad (5.9)$$

where  $z_i$  is defined as before. (24:17)

### Anderson-Darling

The Anderson-Darling distance measure is the measure used for this research. It is a special case of the Cramer-von Mises distance measure where

$$\psi(F(x)) = 1/F(x)(1-F(x)) \quad (5.10)$$

This weighting factor causes more emphasis to be placed upon the tails of the distribution. (24:18) The statistic can be computed by

$$A^2 = \left[ - \left( \sum_{i=1}^n (2i-1) (\ln(z_i) + \ln(1-z_{n+1-i})) \right) / n \right] - n \quad (5.11)$$

where  $z_i$  is defined as before. (1:765) For the mixed normal

$$A^2 = \left[ - \left[ \sum_{i=1}^n (2i-1) \left( \ln \left( p \int_{-\infty}^{x_i} (2\pi\sigma_1^2)^{-1/2} \exp \left[ -1/2(y - \mu_1/\sigma_1)^2 \right] dy \right. \right. \right. \right. \\ + (1-p) \int_{-\infty}^{x_i} (2\pi\sigma_2^2)^{-1/2} \exp \left[ -1/2(y - \mu_2/\sigma_2)^2 \right] dy \Big) \\ + \ln(1 - (p \int_{-\infty}^{x_{n+1-i}} (2\pi\sigma_1^2)^{-1/2} \exp \left[ -1/2(y - \mu_1/\sigma_1)^2 \right] dy \\ + (1-p) \int_{-\infty}^{x_{n+1-i}} (2\pi\sigma_2^2)^{-1/2} \exp \left[ -1/2(y - \mu_2/\sigma_2)^2 \right] dy \Big) \Big) \Big] / n \right] \\ - n \quad (5.12)$$

and for the mixed exponential

$$A^2 = \left[ - \left[ \sum (2i-1) \left( \ln(p(1-\exp(-x_i/\theta_1)) + (1-p)(1-\exp(-x_i/\theta_2))) \right. \right. \right. \\ + \ln(1 - (p(1-\exp(-x_{n+1-i}/\theta_1)) + (1-p)(1-\exp(-x_{n+1-i}/\theta_2)))) \Big) \Big] \\ / n \Big] - n \quad (5.13)$$

These are the equations which will be minimized in this research to obtain the minimum distance estimate of  $p$  where all other parameters have been previously estimated and are assumed true.

## VI. Methodology

The methodology developed in this thesis assumes that the underlying distributions are mixtures of distributions from the families of the Normal distributions and the One-Parameter Exponential distributions, as defined in equations 4.6 and 4.20 respectively. In 1984, Woodward, Parr and associates applied Minimum Distance techniques to the mixed normal, the mixed double exponential, and the mixed t-distribution.(27) In Parr's research, the distance measure chosen was the Cramer-von Mises goodness-of-fit statistic which differs from the current research. However, this research is an extension of Parr's study.

### Monte Carlo Simulation

Monte Carlo simulation techniques have been widely used in previous parameter estimation studies. The technique allows you to specify a particular underlying distribution from which to draw the sample and, in doing so, allows the performance of the estimators to be checked. The Monte Carlo simulation applied to this research consists of the following steps.

- (1) Generate the data from the specified distributions.
- (2) Determine the initial parameter estimates either by using the Method of Moments or quasi-clustering



techniques.

(3) Determine the Minimum Distance estimates of  $p$ , using the estimates of step two as starting values.

(4) Compare the performance of the estimators.

Each of these steps will be explained in detail in the remainder of this chapter. At this point, it should be stated that the computer time required to accomplish this simulation was very great. This limited the number of cases that could be investigated.

#### Sample Generation from the Mixed Normal

Six samples from the mixed normal distribution were generated using the method described by Parr and associates (27). The first step was to generate 100 standard normal deviates by use of the IMSL subroutine GGNPM(23). The second step in the process was to determine which component of the mixture the deviate would be assigned to. This was done by first selecting the desired mixing proportion, then generating 100 (0,1) uniform deviates from the IMSL subroutine GGUBS (23). The normal deviates were assigned to the first component if the uniform deviate was less than the mixing proportion,  $p$ , and to the second if it was greater than  $p$ . The final step was to scale the normal deviates according to the assigned component. An observation from  $f_1(x) = Y$  where

$Y$  was the standard normal deviate, and an observation from  $f_2(x) = Y+b$  where  $b>0$  and is defined as the difference in location parameters. The parameter combinations used were

$$p=.75 \quad \mu_1=0.0 \quad \mu_2=2.32 \quad \sigma_1 = \sigma_2=1$$

$$p=.75 \quad \mu_1=0.0 \quad \mu_2=3.6 \quad \sigma_1 = \sigma_2=1$$

$$p=.50 \quad \mu_1=0.0 \quad \mu_2=2.56 \quad \sigma_1 = \sigma_2=1$$

$$p=.50 \quad \mu_1=0.0 \quad \mu_2=3.76 \quad \sigma_1 = \sigma_2=1$$

$$p=.25 \quad \mu_1=0.0 \quad \mu_2=2.32 \quad \sigma_1 = \sigma_2=1$$

$$p=.25 \quad \mu_1=0.0 \quad \mu_2=3.6 \quad \sigma_1 = \sigma_2=1$$

These combinations were selected so that the analysis of this study could be compared to that of Parr (27).

#### Sample Generation from the Mixed Exponential

Sample generation for the two component mixture of one parameter exponentials was much simpler than that of the mixed normal. Nine samples of size 750 were generated using the IMSL subroutine GGEXT(23). This subroutine automatically generates deviates from a specified mixture and therefore requires no manipulation. The generated samples had the following parameter combinations.

$p=.75$        $\theta_1=2.0$        $\theta_2=0.5$

$p=.50$        $\theta_1=2.0$        $\theta_2=0.5$

$p=.25$        $\theta_1=2.0$        $\theta_2=0.5$

$p=.75$        $\theta_1=1.0$        $\theta_2=0.5$

$p=.50$        $\theta_1=1.0$        $\theta_2=0.5$

$p=.25$        $\theta_1=1.0$        $\theta_2=0.5$

$p=.75$        $\theta_1=3.0$        $\theta_2=0.5$

$p=.50$        $\theta_1=3.0$        $\theta_2=0.5$

$p=.25$        $\theta_1=3.0$        $\theta_2=0.5$

These parameter values were chosen to give varying degrees of separation between the two components.

After the samples were generated the initial parameter estimates had to be determined. This turned out to be a very difficult task.

#### Initial Parameter Estimates for the Normal Mixture

Due to the complexity of both the Method of Moments equations(4.13-4.17) and the Maximum Likelihood equations

(4.35-4.37), the solutions to these equations are both difficult and time consuming to obtain. Because of this an alternate method was chosen. In Woodward and Parr's paper (27), they obtained starting values for the ML technique from an easy to implement quasi-clustering technique. In their final analysis these initial estimates were found to be highly competitive with the maximum likelihood estimates. For this reason and because of its simplicity, this technique was selected for use in this research and it is explained below.

First  $\mu_1 \leq \mu_2$  is assumed to be true. Next, the initial estimate of the mixing proportion,  $p_0$ , is allowed to take on the values of .1,.2,....,9 only. The sample is then ordered and divided into two subsamples,  $Y_1, Y_2, \dots, Y_n$  and  $Y_{n+1}, \dots, Y_n$ , for each allowed  $p_0$  value, where  $n_1 = np$  rounded to the nearest integer. The initial estimate of  $p$  is the value  $p_0$  that maximizes

$$p(1-p)(m_1 - m_2)^2 \quad (6.1)$$

where  $m_j$  is the sample median of the  $j$ th subsample.

The choice for the estimates of  $\mu_1$  and  $\mu_2$  are the corresponding  $m_1$  and  $m_2$  values of the above  $p$  estimate. Finally, the estimates of  $\sigma_1^2$  and  $\sigma_2^2$  are determined by the following equations.

$$\sigma_1^2(0) = ((m_1 - r_1(.25)) / .6745)^2 \quad (6.2)$$

$$\theta_2^2(0) = ((r_2^{(.75)} - m_2) / .6745)^2 \quad (6.3)$$

where  $r_j^{(q)}$  is the qth quartile from the jth cluster.

#### Initial Parameter Estimates for the Exponential Mixture

The Method of Moments technique was chosen to obtain the starting estimates for the case of the mixed exponential. The moment equations were defined in Chapter 4, equations (4.25, 4.26, 4.27). The first attempt to solve these equations was made using the method of false position. These attempts failed however. Instead of converging they diverged and therefore a new method was searched for. The method selected was that of Paul Rider(34).

Rider suggested combining the three moment equations. In doing so and then simplifying, the following quadratic is obtained

$$6(2m'_1{}^2 - m'_2) \theta_j^{*2} + 2(m'_3 - 3m'_1 m'_2) \theta_j^* + 3m'_2{}^2 - 2m'_1 m'_3 = 0 \quad (6.4)$$

where  $m'_i$  is the ith sample moment. (34:144)

The roots this quadratic are estimates of  $\theta_1$  and  $\theta_2$  and in this research they were obtained by the use of the IMSL sub-routine ZQADR.(23) With these two estimates an estimate of p can be obtained from

$$p^* = (m'_1 - \theta_2^*) / (\theta_1^* - \theta_2^*) \quad (6.5)$$

### Minimum Distance Estimates

After obtaining the initial estimates, the minimum distance estimates of the mixing proportion were found for both the normal and the exponential mixture. Refer to Chapter 5 for the explanation of minimum distance methods.

The distance measure selected for this research was the Anderson-Darling goodness-of-fit statistic. In other words, the parameter estimates were selected so as to minimize this statistic. The method used to perform this minimization was the Golden Search. The following is a general algorithm for the method. (28)

Assumption:  $f(x)$  is unimodal on  $[a,b]$  and  $r = .618034$

Step 0. Let  $\underline{x}=a$  and  $\bar{x}=b$

$$x_L = \bar{x} - r(\bar{x} - \underline{x})$$

$$x_R = \underline{x} + r(\bar{x} - \underline{x})$$

Step 1. Evaluate  $f(x_L)$  and  $f(x_R)$

Step 2.  $f(x_L) \geq f(x_R) \Rightarrow \bar{x} = x_R ; x_R = x_L$

$$x_L = \bar{x} - r(\bar{x} - \underline{x})$$

$$f(x_L) \leq f(x_R) \Rightarrow \underline{x} = x_L ; x_L = x_R$$

$$x_R = \underline{x} + r(\bar{x} - \underline{x})$$

$$L = \bar{x} - \underline{x}$$

Step 3. If  $L < \epsilon$ , stop  $\underline{x} \leq x^* \leq \bar{x}$

Otherwise, go to step 1.

For this research  $f(x)$ , the Anderson-Darling statistic, is defined in equation 5.12 for the mixed normal and in equation 5.13 for the mixed exponential.  $\epsilon$  was chosen to be .0001, and the initial values of  $a$  and  $b$  are  $D$  and 1 respectively, since  $0 < p < 1$ .

In this research, when evaluating  $f(x)$  the c.d.f. for the appropriate distribution must also be evaluated. In the case of the mixed normal however, the c.d.f. must be approximated. This approximation was performed by the IMSL subroutine MDNOR.(23)

### Comparison Criteria

After both parameter estimations have been found, the process is repeated 500 times. This is done for both the mixed normal and the mixed exponential. Using the estimates from the 500 replications, the mean square error for both mixing proportion estimates was calculated by

$$MSE = 1/500 \sum_{i=1}^{500} (p_i - p)^2 \quad (6.6)$$

A comparison of the two different MSE calculations gave the indication of which estimation was best.

A sample of the computer code used to implement the above simulation for the mixed exponential can be found in the appendix. It should be noted once again that computer time needed for the program was very large when run on AFIT's SSC computer system.



## VI. Conclusions and Recommendations

### Summary

This thesis compares the Method of Moments and quasi-clustering estimation techniques to Minimum Distance techniques when used to estimate the mixing proportions of mixed one parameter exponentials and mixed normals, respectively. The Method of Moments estimates were used as the baseline for the mixed exponentials, while the quasi-clustering estimates were the baseline for the mixed normals. This comparison was carried out by a computer routine which performed a Monte Carlo simulation, then calculated the mean square errors. These errors were used as the criteria for comparison.

### Conclusions

At the beginning of this research it was hoped that the minimum distance estimation technique would provide better estimates than the baseline techniques. In the case of the mixed exponential this does not appear to be true as can be seen in table I. However, the small mean square errors show both estimators to be good and despite the fact that the Method of Moments technique appears better, it has problems.

First of all, the method of moments technique requires very large sample sizes in order to ensure that the roots of equation 6.4 are positive. (32:144) Even in the case of the sample size of 500 used for this research, one of the para-

meter combinations still became infeasible. (Denoted on table 1 by \*). On the other hand, the minimum distance technique will converge in relatively few iterations with any decent initial estimates, ie. educated guesses, graphical estimates, etc. Secondly, the method is much more difficult to apply than the minimum distance routine due to the complicated moment calculations. For this reason, and due to the closeness of the mean square errors, the method of minimum distance is recommended.

In the case of the mixed normal, the results were much different. Parr's analysis, which used the Cramer-von Mises statistic, had results indicating similar mean square errors for the minimum distance and the clustering estimation techniques. In the current research however, the clustering method proved to be superior. This was due to the inability of the Anderson-Darling statistic to distinguish between the .75 and the .25 mixing proportions. For this reason, use of the Anderson-Darling statistic as a distance measure for the minimum distance technique, when applied to mixed normals, is not recommended.

#### Suggested Research

Extending the above research to include mixed Weibulls is suggested. Due to the strength and versatility of the Weibull distribution as a failure model, this extension could be a powerful aid in the search for better parameter estimation techniques.

TABLE I.  
Simulation Results for Mixtures of  
Two Exponential Components

| True Mean<br>Values            | Mixing<br>Proportion | Estimator | MSE       |
|--------------------------------|----------------------|-----------|-----------|
| $\mu_1 = 2.0$<br>$\mu_2 = 0.5$ | .75                  | MDE       | .0428782  |
|                                |                      | Moment    | .0339208  |
|                                | .50                  | MDE       | .0281973  |
|                                |                      | Moment    | .0208788  |
|                                | .25                  | MDE       | .0159149  |
|                                |                      | Moment    | .0134047  |
| $\mu_1 = 1.0$<br>$\mu_2 = 0.5$ | .75                  | MDE       | *         |
|                                |                      | Moment    |           |
|                                | .50                  | MDE       | .0910827  |
|                                |                      | Moment    | .0862087  |
|                                | .25                  | MDE       | .105535   |
|                                |                      | Moment    | .101567   |
| $\mu_1 = 3.0$<br>$\mu_2 = 0.5$ | .75                  | MDE       | .0381142  |
|                                |                      | Moment    | .0235877  |
|                                | .50                  | MDE       | .0265675  |
|                                |                      | Moment    | .0136226  |
|                                | .25                  | MDE       | .011018   |
|                                |                      | Moment    | .00732790 |

TABLE II.  
Simulation Results for Mixtures of  
Two Normal Components  
(equal shape parameters)

| True Mean<br>Values | Mixing<br>Proportion | Estimator         | MSE                   |
|---------------------|----------------------|-------------------|-----------------------|
| 1 = 0.0<br>2 = 2.32 | .75                  | MDE<br>Clustering | .122337<br>.0183400   |
| 1 = 0.0<br>2 = 2.56 | .50                  | MDE<br>Clustering | .0338620<br>.0114600  |
| 1 = 0.0<br>2 = 2.32 | .25                  | MDE<br>Clustering | .332602<br>.0194600   |
| 1 = 0.0<br>2 = 3.6  | .75                  | MDE<br>Clustering | .153062<br>.00726004  |
| 1 = 0.0<br>2 = 3.76 | .50                  | MDE<br>Clustering | .0221589<br>.00671999 |
| 1 = 0.0<br>2 = 3.6  | .25                  | MDE<br>Clustering | .309898<br>.00826004  |

# APPENDIX: COMPUTER CODE

```
C *****
C
C      This program generates estimates of the mixing pro-
C      portion fo two exponential distributions using the
C      both the Method of Moments and the Minimum Distance
C      estimation techniques.
C
C *****

real samp(750),true(3),temp,min(3),mse,a,b,c
real sammom(3),mean1,mean2,estip,x,y,z,msemom,temp(3)
real x1,y1,z1
integer nr,n,m,d,ier
complex zsm,zlg
common true(3)
double precision dseed
external ggext

nr=750
true(1)= 2.0
true(2)= .50
true(3)= .25
mse=0.0
msemom=0.0
print*, ' '
print*, 'exponential: mean1=2.0,mean2=.50,p=.25'
print*, ' '
dseed=295847397.d0

C *****
C
C      Begin the iteration loop
C
C *****

do 10 k=1,500
sammom(1)=0.0
sammom(2)=0.0
sammom(3)=0.0
p=.25
true(1)=2.0
true(2)=.50
true(3)=p
xm1=2
xm2=.5
nr=750

C *****
C
C      Generate deviates from a mixture of two exponentials
C      by the use of the IMSL routine GGTEXT and sort the
```

```

C      data.
C
C      *****
call ggtext(dseed,p,xm1,xm2,nr,samp,ier)
print*, ' '
do 500 i=1,749
    d=i+1
    do 400 j=d,750

        if (samp(i) .gt. samp(j)) then
            temp=samp(i)
            samp(i)=samp(j)
            samp(j)=temp

        end if

400    continue

500    continue

C      *****
C      Calculate the first three sample moments.
C      *****

do 600 i=1,750
    sammom(1)=sammom(1)+samp(i)
    sammom(2)=sammom(2)+samp(i)**2
    sammom(3)=sammom(3)+samp(i)**3
600    continue
    sammom(1)=sammom(1)/750
    sammom(2)=sammom(2)/750
    sammom(3)=sammom(3)/750

C      *****
C      Find the moment estimators by use of the IMSL rou-
C      tine ZQADR. This routine finds the roots of a qua-
C      dratic.
C      *****

print*, ' '
nr=750
m=3
n=2
x1=sammom(1)
y1=sammom(2)
z1=sammom(3)
x=x1
y=y1
z=z1
a=6*(2*x**2-y)

```

```

b=2*(z-3*x*y)

c=3*y**2-2*x*z
call zqadr(a,b,c,zsm,zlg,ier)
print*, 'sammom check', (sammom(i), i=1,3)
mean1=real(zlg)
mean2=real(zsm)
estip=(sammom(1)-mean2)/(mean1-mean2)
print*, 'method of moments estimates', mean1, mean2, estip

C *****
C
C Calculate the minimum distance estimates by using
C the Golden Search method to minimize the Anderson-
C Darling Goodness-of-fit statistic.
C *****

tempt(1)=true(1)
tempt(2)=true(2)
tempt(3)=true(3)
true(1)=mean1
true(2)=mean2
true(3)=estip
call mindis(nr,samp,m,min)
print*, ' '
print*, 'minimum distance estimates'
print*, 'iteration ', k
print*, (min(i), i=1,3)

C *****
C
C Calculate the mean square error for the minimum dis-
C tance and the method of moments estimators.
C *****

mse=mse+(min(3)-p)**2
msemom=msemom+(estip-p)**2
10 continue
mse=mse/500
msemom=msemom/500

print*, ' '
print*, 'mse= ', mse
print*, ' '
print*, 'msemom= ', msemom
stop
end

C *****
C

```

```

c      Subroutine gof. Calculates the Anderson-Darling gof
c      statistic for the Golden Search routine.

```

```

c
c      *****

```

```

      subroutine gof(nr,samp,m,adgof)
      real samp(750),true(3),adgof,cum1(750),cum2(750),z(750)
      real sum,temp
      common true(3)
      integer nr,m,i
      nr=750

      do 10 i=1,nr
        cum1(i)=1-exp(-1*samp(i)/true(1))
        cum2(i)=1-exp(-1*samp(i)/true(2))
        z(i)=true(3)*cum1(i)+(1-true(3))*cum2(i)
        if (z(i).lt.0.001)then
          z(i)=0.001
        end if
        if (z(i).gt.0.000)then
          z(i)=.999
        end if
10    continue
      sum=0.0
      do 20 i=1,nr
        temp=(2*i-1)*(alog(z(i))+alog(1-z(nr+1-i)))
        sum=sum+temp
20    continue
      adgof=((-1*sum)/nr)-nr
      return
      end

```

```

c      *****

```

```

c      Subroutine mindis. Calculates the minimum distance
c      estimates by means of a golden search

```

```

c      *****

```

```

      subroutine mindis(nr,samp,m,min)
      real samp(750),true(3),min(3),nr,m,adgof,fa,fb,step,r,a
      real b,left,right,fleft,fright
      integer i
      common true(3)

      step=.01
      r=.618034
      a=0.0
      true(3)=0.0
      call gof(nr,samp,m,adgof)

```



```

fa=adgof
b=1.0
true(3)=left

call gof(nr,samp,m,adgof)
fleft=adgof
true(3)=right
call gof(nr,samp,m,adgof)
fright=adgof
i=0
40 continue
if(fleft.lt.fright) then
    b=right
    fb=fright
    right=left
    fright=fleft
    left=b-r*(b-a)
    true(3)=left
    call gof(nr,samp,m,adgof)
    fleft=adgof
else
    a=left
    fa=fleft
    left=right
    fleft=fright
    right=a+r*(b-a)
    true(3)=right
    call gof(nr,samp,m,adgof)
    fright=adgof
end if
i=i+1
if(abs(fb-fa).gt..0001.and.i.lt.200)then
    go to 40
else
    go to 50
end if
50 continue

min(1)=true(1)
min(2)=true(2)
min(3)=true(3)
print*,' '
print*,'adgof'
print*,adgof
return
end

```

### Bibliography

1. Anderson, T.W. and Darling, O.A. "A Test of Goodness of Fit," American Statistical Association Journal, :765-769 (Dec 1954).
2. Ashour, S.K. and Jones, P.W. "Bayesian Estimation of the Parameters of the Mixed Exponential Distribution from Censored Samples," Biom. Z. Bd., 18:633-637 (1976)  
18:633-637 (1976)
3. Banks, Jerry and Carson, John S. II. Discrete - Event System Simulation. New Jersey :Prentice-Hall, Inc., 1984.
4. Basu, J.P. and Odell, P.L. "Concerning Several Methods for Estimating Crop Acreages Using Remote Sensing Data," Communications in Statistics - Theory and Methods, A5: 1091-1114 (1976).
5. Boardman, J.J. On the Maximum Likelihood Estimation of the Parameters of Mixed Exponential Distributions with Applications to Life Testing, PhD thesis , Rutgers - The State University, New Jersey, 1968.
6. Bowden, D.C. and Dick, N.P. "Maximum Likelihood Estimation for Mixtures of Two Normal Distributions," Biometrics, 29:781-790 (1973).
7. Chang, W.C. and Tan, W.Y. "Some Comparisons of the Method of Moments and the Method of Two Normal Distributions," Journal of the American Statistical Association, 67:702-708 (1972).
8. Cohen, A. Clifford, Jr. "Estimating in Mixtures of Poisson and Mixtures of Exponential Distributions," NASA Technical Memorandum, NASA TM X-53245, George C. Marshall Space Flight Center, Huntsville, Alabama (1965).
9. Cohen, A. Clifford, Jr. "Estimation in Mixtures of Two Normal Distributions," Technometrics, 9:15-28 (Feb 1967).
10. Daniels, Tony G. Robust Estimation of the Generalized t Distribution Using Minimum Distance Estimation. MS thesis. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB, Ohio, December 1980.
11. Day, N.E. "Estimating the Components of a Mixture of Two Normal Distributions," Biometrika, 56:463-474 (1969).
12. Dick, N.P. and Hosmer, D.W. "Information and Mixture of Two Normal Distributions," Journal of Statistical Computation and Simulation, 6:137-148 (1977).

13. Do, K. and McLachlan, G.J. "Estimation of Mixing Proportions: A Case Study,"
14. Falls, L.W. "Estimation of Parameters in Parameters in Compound Weibull Distributions," Technometrics, 12:399-407 (May 1970).
15. Hader, J. R. and Mendenhall, W. "Estimation of Parameters of Mixed Exponentially Failure Time Distributions from Censored Life Test Data," Biometrika, 45:504-520 (1958).
16. Hasselblad, V.A. "Estimation of Parameters for a Mixture of Normal Distributions," Technometrics, 8:431-446 (1966).
17. Hasselblad, V.A. "Estimation of Finite Mixtures of Distributions From the Exponential Family," Journal of the American Statistical Association, 64:1459-1471 (1969).
18. Hill, B.M. "Information for Estimating the Proportions in Mixture of Exponential and Normal Distributions," Journal of the American Statistical Association, 56:463-474 (1963).
19. Hobbs, Jon R., James, William, and Moore, Albert H. "Minimum - Distance Estimation of the Three Parameters of the Gamma Distribution," IEEE Transactions on Reliability, R-33, No.3:237-240 (1984).
20. Hobbs, Jon R., Miller, Robert M. and Moore, Albert H. "Minimum - Distance Estimation of the Parameters of the 3-Parameter Weibull Distribution," IEEE Transactions on Reliability, R-34, No 5:495-496 (Dec 1985).
21. Hosmer, D.W. "A Comparison of Iterative Maximum Likelihood Estimates of the Parameters of a Mixture of Two Normal Distributions Under Three Different Types of Samples," Biometrics, 29:761-770 (1973).
22. Hosmer, D.W. "On MLE of the Parameters of a Mixture of Two Normal Distributions," Communications in Statistics, 1:217-227 (1973).
23. IMSL Inc. IMSL Library Reference Manual. Houston: IMSL Inc., 1980.
24. James, Capt. William L. Robust Minimum Distance Estimation Based on a Family of Three - Parameter Gamma Distributions. MS thesis. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 1980.
25. James, I.R. "Estimation of the Mixing Proportion in a Mixture of Two Normal Distributions from Simple, Rapid

- Measurements," Biometrics, 34:265-275 (June 1978).
26. Kao, J.H.K. "A Graphical Estimation of Mixed Weibull Parameters in Life - Testing of Electron Tubes," Technometrics, 1:389-407 (Nov 1959).
  27. Lindsey, Hildegard, Parr, William C., Schucany, William R., Woodward, Wayne A. "A Comparison of Minimum Distance and Maximum Likelihood Estimation of a Mixture Proportion," Journal of the American Statistical Association, 79:590-598 (Sept 1984).
  28. Litko, Capt Joseph R. Lecture materials distributed in OPER 663, Deterministic Methods in Operations Research. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB, OH, (Nov. 1985).
  29. Mann, Nancy R., Schafer, Ray. E. and Singpurwalla, Nozer D. Methods for Statistical Analysis of Reliability and Life Data. John Wiley and Sons, Inc., 1974.
  30. Mendenhall, William, Scheaffer, Richard L. and Wackerly, Dennis D. Mathematical Statistics with Applications. Boston : Duxbury Press, 1981.
  31. Miller, Robert M. Robust Minimum Distance Estimation of the Three Parameter Weibull Distribution. MS thesis. School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB, Ohio, December (1980).
  32. Odell, P.L. "Current and Past Roles of the Statistician in Space Applications," Communications in Statistics - Theory and Methods, A5:1077-1089 (1976).
  33. Parr, William C. and Schucany, William R. "Minimum Distance and Robust Estimation," Journal of the American Statistical Association, 75:616-624 (1980).
  34. Rider, P.R. "The Method of Moments Applied to a Mixture of Two Exponential Distributions," Annals of Mathematics and Statistics, 32:143-147 (1961).
  35. Rider, Paul R. "Estimating the Parameters of Mixed Poisson, Binomial, and Weibull Distributions by the Method of Moments," Bulletin de l'Institut International de Statistique, 39 (1962).
  36. Wolfowitz, J. "Estimation by the Minimum Distance Method," Annals of the Institute of Statistical Mathematics (Japan), 5:9-23 (1953).
  37. Wolfowitz, J. "The Minimum Distance Method," Annals of Mathematical Statistics, 28:75-88 (1957).

### Vita

Lieutenant R. Nicole Benton-Santo was born on 24 November 1963 in Stephenville, Newfoundland, Canada. She graduated from Douglas High School at Ellsworth AFB, South Dakota in 1981. She attended Creighton University in Omaha, Nebraska from which she received the degree of Bachelor of Science in Mathematics in May 1985. Upon graduation, she received a commission in the USAF through the ROTC program, immediately entered active duty and reported to the School of Engineering, Air Force Institute of Technology.

Permanent address: P.O. Box 953

Kannapolis, N. Carolina

28081

| REPORT DOCUMENTATION PAGE   |       |  |  | Form Approved<br>OMB No. 0704-0188                     |             |
|---|-------|--|--|--|-------------|
| 1a. REPORT SECURITY CLASSIFICATION<br>unclassified  |       |  | 1b. RESTRICTIVE MARKINGS   |  |             |
| 2a. SECURITY CLASSIFICATION AUTHORITY   |       |  | 3. DISTRIBUTION / AVAILABILITY OF REPORT<br>Approved for public release;<br>distribution unlimited |  |             |
| 2b. DECLASSIFICATION / DOWNGRADING SCHEDULE   |       |  |  |  |             |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S)<br>AFIT/GOR/MA/86D-1  |       |  | 5. MONITORING ORGANIZATION REPORT NUMBER(S)  |  |             |
| 6a. NAME OF PERFORMING ORGANIZATION<br>School of Engineering  |       | 6b. OFFICE SYMBOL<br>(if applicable)<br>AFIT/ENS |  | 7a. NAME OF MONITORING ORGANIZATION                    |             |
| 6c. ADDRESS (City, State, and ZIP Code)<br>Air Force Institute of Technology<br>Wright-Patterson AFB, Ohio 45433  |       |  |  | 7b. ADDRESS (City, State, and ZIP Code)                |             |
| 8a. NAME OF FUNDING / SPONSORING ORGANIZATION   |       | 8b. OFFICE SYMBOL<br>(if applicable)             |  | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER        |             |
| 8c. ADDRESS (City, State, and ZIP Code)   |       |  |  | 10. SOURCE OF FUNDING NUMBERS                          |             |
|   |       | PROGRAM<br>ELEMENT NO.                           |  | PROJECT<br>NO.   | TASK<br>NO. |
|   |       |  |  | WORK UNIT<br>ACCESSION NO.                             |             |
| 11. TITLE (Include Security Classification)<br>MINIMUM DISTANCE ESTIMATION OF MIXTURE PROPORTIONS   |       |  |  |  |             |
| 12. PERSONAL AUTHOR(S)<br>Robin Nicole Benton-Santo, B.S.Mth., 2dLt., USAF  |       |  |  |  |             |
| 13a. TYPE OF REPORT<br>MS Thesis  |       | 13b. TIME COVERED<br>FROM _____ TO _____         |  | 14. DATE OF REPORT (Year, Month, Day)<br>1986 December |             |
|   |       |  |  | 15. PAGE COUNT<br>53                                   |             |
| 16. SUPPLEMENTARY NOTATION  |       |  |  |  |             |
| 17. COSATI CODES  |       |  | 18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)                  |  |             |
| FIELD   | GROUP | SUB-GROUP  |  |  |             |
| 12  | 01    |  | Statistical Analysis, Monte Carlo method,<br>Estimating, Mixed Distributions                       |  |             |
| 19. ABSTRACT (Continue on reverse if necessary and identify by block number)  |       |  |  |  |             |
| Thesis Advisor : Albert H. Moore, PhD.<br>Professor of Mathematics  |       |  |  |  |             |
| Approved for public release: IAW AFR 180-15.<br>LYNN E. WOLAVEN 5 May 87<br>Dean for Research and Professional Development<br>Air Force Institute of Technology (AFIT)<br>Wright-Patterson AFB OH 45433 |       |  |  |  |             |
| 20. DISTRIBUTION / AVAILABILITY OF ABSTRACT<br><input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS                     |       |  |  | 21. ABSTRACT SECURITY CLASSIFICATION<br>unclassified   |             |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL<br>Albert H. Moore, PhD.  |       |  |  | 22b. TELEPHONE (Include Area Code)<br>(513)255-3098    |             |
|   |       |  |  | 22c. OFFICE SYMBOL<br>AFIT/ENC                         |             |

Minimum Distance estimation was used to calculate estimates of the mixing proportion of the mixture of two normal distributions and the mixture of two exponential distributions. The estimation was carried out by using the Golden Search technique to minimize the Anderson-Darling goodness-of-fit statistic. A Monte Carlo simulation was run for both distribution mixtures, varying the mixing proportions from .25, .5 to .75 with sample sizes of 100 for the normal mixture and 750 for the mixture of exponentials. The simulation was run 500 times for each parameter combination.

An ad hoc quasi-clustering technique was used to obtain the initial estimates for the parameters of the mixed normal while the method of moments technique was used to obtain initial estimates for the mixed exponential parameters. These estimates were then used to start the minimum distance routines which were used to obtain new estimates of the mixing proportions.

Finally, the mean square errors were calculated for use as a means of comparison for the different estimation procedures.

